Evidence for strong mutation bias towards, and selection against, U content in SARS-CoV-2: implications for vaccine design

Evidence for strong mutation bias towards, and selection against, U content in SARS-CoV-2: implications for vaccine design.

Alan M. Rice*[1], Atahualpa Castillo Morales*[1], Alexander T. Ho*[1], Christine Mordstein[1,2], Stefanie Mühlhausen[1], Samir Watson[3], Laura Cano[2], Bethan Young[1,2], Grzegorz Kudla[2§] and Laurence D. Hurst[1§]

1. The Milner Centre for Evolution, Department of Biology and Biochemistry, University of Bath, Bath, BA2 7AY, UK
2. MRC Human Genetics Unit, Institute for Genetics and Molecular Medicine, The University of Edinburgh, Edinburgh, UK
3. Aarhus University, Department of Molecular Biology and Genetics, C F Møllers Allé 3, 8000 Aarhus, Denmark.

*These authors contributed equally
§Co-senior authors

ABSTRACT

Large-scale re-engineering of synonymous sites is a promising strategy to generate vaccines either through synthesis of attenuated viruses or via codon optimized genes in DNA vaccines. Attenuation typically relies on de-optimisation of codon pairs and maximization of CpG dinucleotide frequencies. So as to formulate evolutionarily-informed attenuation strategies that aim to force nucleotide usage against the direction favoured by selection, here we examine available whole-genome sequences of SARS-CoV-2 to infer patterns of mutation and selection on synonymous sites. Analysis of mutational profiles indicates a strong mutation bias towards U. In turn, analysis of observed synonymous site composition implicates selection against U. Accounting for dinucleotide effects reinforces this conclusion, observed UU content being a quarter of that expected under neutrality. Possible mechanisms of selection against U mutations includes selection for higher expression, for high mRNA stability or lower immunogenicity of viral genes. Consistent with gene-specific selection against CpG dinucleotides, we observe systematic differences of CpG content between SARS-CoV-2 genes. We propose an evolutionarily-informed approach to attenuation that, unusually, seeks to increase usage of the already most common synonymous codons. Comparable analysis of H1N1 and Ebola finds that GC3 deviated from neutral equilibrium is not a universal feature, cautioning against generalization of results.

Multiple strategies towards the development of a SARS-CoV-2 vaccine are being pursued (Thanh Le, et al. 2020). These include attenuated or inactivated viruses, replicating and non-replicating viral vectors, proteins, and nucleic acids (reviewed in (Thanh Le, et al. 2020)). Some of these strategies, notably DNA or RNA vaccines, rely on the expression of viral genes in humans. These and other modes of vaccine development (e.g. to produce high protein titres) might benefit from synonymous site modification (Gustafsson, et al. 2004; Coleman, et al. 2008; Kudla, et al. 2009; Fath, et al. 2011; Bentele, et al. 2013; Mordstein, et al. 2020). Coding sequence optimization methods can be directed to modification of codon usage, codon pair usage, nucleotide and dinucleotide content, and other properties of coding sequences, with the aim of achieving a desired phenotype, such as increased gene expression (Gustafsson, et al. 2004; Kudla, et al. 2009; Fath, et al. 2011; Bentele, et al. 2013; Mordstein, et al. 2020), improved immunogenicity (Stachyra, et al. 2016) or virus attenuation (Coleman 2008).

The DNA and RNA vaccine design methods that might benefit from synonymous site modification fall broadly into two classes: those that aim to detune the live virus (e.g. Coleman, et al. 2008; Mueller, et al. 2010; Manokaran, et al. 2019; Cai, et al. 2020) and those that aim to enhance expression of individual genes (e.g. Stachyra, et al. 2014). As with the expression of any transgene, if one viral gene alone is to be expressed in a vector, for example as part of a DNA vaccine (Stachyra, et al. 2014; Stachyra, et al. 2016), then codon optimization of the gene concerned to enable high gene expression may be desirable, not least because such genes can improve immunogenicity (Stachyra, et al. 2016), thereby requiring fewer doses (Wang, et al. 2006). Such DNA based vaccines are regarded as relatively safe as no infective form of the virus is required (Khan 2013).

Viral attenuation differs from other coding sequence optimization strategies in that it aims to produce gene sequences with low expression levels, with the assumption that this will lead to the production of intact (or near intact) viruses with low pathogenicity, which can nevertheless induce an immune response in the host (Coleman, et al. 2008). Synthesis of a complete attenuated virus with detuned synonymous sites can however result in a virus almost unable to replicate (Coleman, et al. 2008) and, as such, a mosaic synthetic virus, with some genes deoptimized some not, can be preferable (Coleman, et al. 2008). Attenuation via modification of many synonymous sites has the advantage that any such virus employed as a vaccine will likely need many mutations to acquire wild-type fitness. Such a strategy is thus likely to be robust to virus/vaccine intra-host evolution (Coleman, et al. 2008), this being reinforced by the relatively low mutation rate of SARS-CoV-2 (about 1 mutation every two weeks, 26.9 per year (Hill and Rambaut 2020; Nextstrain 2020)). Synonymous codon manipulation has thus been proposed as a viable strategy for SARS-CoV-2 attenuation and vaccine production (e.g. Kames, et al. 2020). A live attenuated codon deoptimized vaccine is being attempted by three groups (as of July 14th report of World Health Organization 2020). For further consideration of development and safety aspects of SARS-CoV-2 vaccines see Peeples (2020) and Amanat and Krammer (2020).

Viral attenuation can be achieved by alteration of synonymous sites as a means to modify the pattern of dinucleotides that bridge between successive codons (alias codon pair bias) while retaining the original protein (Karlin, et al. 1994; Rima and McFerran 1997; Coleman, et al. 2008). This codon pair bias attenuation effect has recently been shown to be largely owing to increased CpG content (Tulloch, et al. 2014; Gaunt, et al. 2016). This is very likely to relate to the activity of the human Zinc Antiviral Protein (ZAP) as this targets transcripts with high CpG content (Takata,

et al. 2017; Ficarelli, et al. 2020), although it is by no means the only antiviral protein (Supplementary Table 1). As might be expected, ZAP is under positive selection owing to host-parasite coevolution (Kerns, et al. 2008). This activity of ZAP suggests a simple attenuation strategy for SARS-CoV-2, i.e. to increase CpG content (Kames, et al. 2020), this being consistent with the observed low CpG enrichment of the virus as sequenced in the wild (Xia 2020), also seen in cytoplasmic viruses more generally (Simmonds, et al. 2013). UpA is commonly considered alongside CpG not least because both are under-represented in native human transcripts (Simmonds, et al. 2013) and UpA is cleaved by RNAseL (Odon, et al. 2019). Similarly, viruses lacking CpG also tend not to have UpA and engineering increased CpG and UpA attenuates viruses (Simmonds, et al. 2013; Odon, et al. 2019). UpA depletion in SARS-CoV-2 is weaker than CpG depletion (see below).

Although codon pair bias and dinucleotide composition have been commonly discussed in the context of virus attenuation, these are not the only coding sequence modification strategies that can conceivably produce attenuated viruses. Recently, codon bias (Radhakrishnan, et al. 2016; Wu, et al. 2019; Buschauer, et al. 2020), nucleotide composition (Kudla, et al. 2006; Mordstein, et al. 2020) and RNA structure (Mauger, et al. 2019) have (re-)emerged as important inter-related determinants of gene expression in mammalian cells. Additionally, viral nucleotide and dinucleotide composition have a known role in the immunogenicity of nucleic acids via TLR-7 (Diebold, et al. 2004). As a result, understanding forces that operate on synonymous site composition, and on nucleotide content more generally, are central to evolutionarily-informed vaccine design, and to our understanding of the biology of SARS-CoV-2. As codon optimization is commonly informed by synonymous site usage in the host genome, we here focus on the relationship between synonymous site selection in the virus and attenuation but are cognisant that lessons learnt may also apply to the optimization problem. Specifically, we aim to discern how selection acts on synonymous sites with a view to engineering the virus against the direction favoured by selection on the virus.

One means to test for selection, or more generally forces causing a fixation bias, is to identify a difference between predicted equilibrium nucleotide composition (or dinucleotide composition) under a neutral-mutation bias model and the values observed in the wild. To perform such a test one requires data on the relative rates of different classes of mutations (A→U, G→C etc) and from these rates per occurrence of the nucleotide calculate the equilibrium position i.e. the nucleotide content at which the rate of gain by mutation from other residues is equal to the rate

of mutational loss. One can then compare observed and neutral equilibrium predicted values, with any discrepancy implicating a fixation bias.

Such methods have revealed commonplace deviations from null neutral expectations. For example, bacteria show a common GC→AT mutational bias (Hershberg and Petrov 2010), and hence a deviation from equilibrium in GC rich bacteria (Hildebrand, et al. 2010). Similarly, non-equilibrium TA nucleotide skews (Charneski, et al. 2011) have been identified. A recent large survey indicated that G+C deviating from neutral equilibria is also common within both prokaryotes and eukaryotes (Long, et al. 2018). To derive this conclusion Lynch and colleagues extracted, from mutation accumulation (MA) experiments or parent-offspring sequencing, mutational profiles for numerous species and showed that the observed G+C content, even at codon third sites, was commonly higher than expected given the profile of mutational events (Long, et al. 2018). The cause of this is unresolved, although GC biased gene conversion is one possible explanation (Long, et al. 2018).

Rapid, accurate and common sequencing of epidemic and pandemic pathogens provide a rich source of data from which to derive the mutational profile (Hershberg and Petrov 2010; Hildebrand, et al. 2010; Charneski, et al. 2011). It is possible to ascribe both ancestral and derived states and hence infer the full mononucleotide mutational matrix (a 4 x 4, 12 parameter matrix of all possible mutations from one state to another) and, with enough mutations, the full dinucleotide matrix (a 16 x 16, 240 parameter matrix of all possible mutations from one dinucleotide to another). Here then we apply this method to SARS-CoV-2.

Under the assumption of selection against CpG (Xia 2020), we predict that observed GC content would be lower than the neutral mutational equilibrium GC content. Under the assumption that synonymous sites are neutrally evolving we expect the predicted equilibrium distribution of the four nucleotides at 4-fold degenerate sites so be no different to that observed. We find in support of neither hypothesis. Our data suggests, unusually, that the most common third site residue (U) is also the one selected against. Given this we thus propose the unusual strategy of increasing the usage of the already most highly used residue so as to degrade performance of the virus. Given that prior evidence indicated that selection for reduced CpG content is particular to just immediate early genes (Lin, et al. 2020), we also propose a "gene-bespoke" approach (i.e. one tailored to each gene's characteristics) sensitive to both CpG and putative selection on synonymous site U.

METHODS

Gene locations

We employed NC_045512 to specify the gene sequence to determine observed GC content, CpG content etc. However, following further annotation of genes (Kim, et al. 2020) we modified the gene locations to reflect those specified: https://github.com/hyeshik/sars-cov-2-transcriptome/blob/master/reference/SARS-CoV-2-annotations.gff. Specifically, to avoid a small codon overlap, we exclude the overlap hence employed annotation:

ORF7a protein 27394..27759 → 27394..27753

ORF7b protein 27756..27887 → 27762..27887

To consider ORF1a and ORF1b independently and to avoid overlap we employ:

ORF1a→ 266-13465

ORF1b → 13471-21552

Estimating flux rates from data

As with parent-offspring sequencing and MA lines, to estimate neutral equilibrium nucleotide content we require that the mutations observed are an unbiased sample of the mutational profile (Hildebrand, et al. 2010; Long, et al. 2018). With very common sequencing (in all cases, short time periods between ancestor and progeny) we can ignore the possibility of multiple sequential hits at the same site (with the first hits going unsequenced) contaminating the mutational matrix. In principle the method can be misled by strong selection purging, in a non-random fashion, mutations prior to their appearing in the population. However, if most selection is weak purifying selection there is then a lag between a deleterious mutation appearing (and being sequenced) and it being purged from a population. Declines in Ka/Ks as time to common ancestry increases in closely related bacteria strains (Rocha, et al. 2006) is consistent with such a model. In principle, even if there is strong selection on some mutations this too need not be problematic, so long as strong selection only affects the observed rate of appearance in sequencing data of new mutations but not the relative proportions of the different mutational classes (C→G, A→U etc). Moreover, if selection does act in a biased manner it should force the predicted equilibrium to more closely resemble the observed nucleotide content, rendering the test conservative. To be cautious, however, we focus on segregating mutations at 4-fold degenerate synonymous sites as the closest approximation to the underlying mutational profile.

15,721 SARS-CoV-2 genome assemblies available on 12/05/020 were downloaded from the GISAID (Shu and McCauley 2017) Initiative EpiCoV platform. Only assemblies flagged as

"complete (>29,000 bp)", "high coverage only", and from a human isolate were downloaded. Isolates with more than 1% of ambiguous base calls (rounded to 298 bases) were removed, leaving 14,855 genomes. Sequences were aligned with MAFFT 7.458 (Katoh and Standley 2013) to Wuhan-Hu-1 reference genome (EPI_ISL_402124). EPI_ISL_402124 was collected from a retailer at Huanan Seafood Wholesale Market, Wuhan on December 30th, 2019. We employed this sequence as not only was it an early sequence, but it also matches the consensus generated from all the 19 sequences that were collected prior to Dec 31st. Variant sites were obtained from the MSA using the package SNP-sites (Page, et al. 2016) and whole genome nucleotide flux estimates were obtained by counting the frequency of each type of mutation with respect to the reference genome. Each given mutation at any given site was counted once, regardless of its frequency within the population. Our method should be insensitive to the presence of recombination, not that there is any evidence that SARS-CoV-2 has recombined through its pandemic phase (Wang, et al. 2020). For consideration of homoplasies (independent mutations at the same site) see below.

Isolates containing at least one coding sequence of length not divisible by three were excluded, removing 58 strains, resulting in a set of 14,599 sequences. CDSs were then translated using BioPython, re-aligned using MAFFT, and then reversed translated using TranslatorX (Abascal, et al. 2010). MSA of CDSs were concatenated and then, just as with the whole genome analysis, variant sites were obtained using SNP-sites and flux estimates were obtained by counting the frequency of each type of change with respect to the reference.

Additionally, H1N1 influenza A pdm09 sequences for strains collected between January 2009 and August 2010 that contained segments PB2, PB1, PA, HA, NP, NA, MP and NS were obtained from GISAID (Shu and McCauley 2017) for 4 segments: RNA polymerase subunit (PB2), hemagglutinin (HA), nucleoprotein (NP), and neuraminidase (NA). Sequences with length not divisible by three or containing a stop codon when translated were excluded. Remaining sequences were translated by BioPython and aligned to Mexican strain EPI_ISL_66702 using MAFFT, and reverse translated to nucleotides using TranslatorX (Abascal, et al. 2010).

Multiple sequence alignment of 1610 full Ebola virus (EBOV) genomes sampled between 17 March 2014 and 24 October 2015 in West Africa was downloaded from EbolaID database (Carneiro and Pereira 2016). The alignment includes the reference genome NC_002549.1. Genomes with a proportion of more than 10% missing sites were discarded. CDSs for each strain

were obtained by extracting the coordinates from the reference genome on the alignment. In order to include in the analysis as the largest proportion of the gene ZEBOVgp4, the longest CDS (NP_066246.1) was used, and the shorter, overlapping proteins NP_066247.1 and NP_066248.1 were discarded. Just as in the case of H1N1, sequences with length not divisible by three were excluded. Remaining sequences were translated aligned to the reference strain using MAFFT, and reverse translated to nucleotides using TranslatorX (Abascal, et al. 2010).

Estimating equilibria

In principle one can estimate neutral GC equilibria knowing relative rates of GC→AT and AT→GC mutations alone e.g. (Long, et al. 2018). However, we take a fuller approach to estimate the equilibrium content of all nucleotides that also enables us to capture nucleotides skews (Charneski, et al. 2011). This has the advantage of treating all four bases as separate independent states, as is fitting for a single stranded virus unconstrained by Chargaff's first parity rule (Elson and Chargaff 1952). Let us denote the frequency of G as G, the frequency of U, U etc. We shall write that the mutational frequency of G to U will be g2u etc, these being measured per occurrence of the starting base. The frequency of the nucleotides after some period (N') will then be:

G' = G (1-g2u-g2c-g2a) + A (a2g) + U (u2g) + C (c2g)

C' = C (1-c2u-c2g-c2a) + A (a2c) + U (u2c) + G (g2c)

A' = A (1-a2u-a2c-a2g) + G (g2a) + U (u2a) + C (c2a)

U' = U (1-u2g-u2c-u2a) + A (a2u) + G (g2u) + C (c2u)

We then solve such that G'=G, U'=U etc. This thus resolves to:

G (g2u+g2c+g2a) = A (a2g) + U (u2g) + C (c2g)

C (c2u+c2g+c2a) = A (a2c) + U (u2c) + G (g2c)

A (a2u+a2c+a2g) = G (g2a) + U (u2a) + C (c2a)

U (u2g+u2c+u2a) = A (a2u) + G (g2u) + C (c2u)

Note that the left hand of each equation is the rate of loss given current abundance, while the right is the rate of gain given current abundances (i.e. we are solving for gain =loss). The 12 flux parameters (a2u, a2c etc) we derive from the mutational profile these being the number of observed changes per relevant occurrence of the nucleotide in the ancestral (pre mutated) sequence. We then solve these four simultaneous equations. Note that we replace any one arbitrarily chosen frequency by 1- sum of the other three (e.g. U = 1-A-C-G). These were solved

in NumPy. Equilibrium solutions we denote with an asterisk (e.g. G*, GC3* etc). N4* implies nucleotide content of nucleotide N at 4-fold degenerate sites.

To assign bounds on the equilibrium estimates we perform a bootstrap test in which we resample with replacement M mutations from the set of M mutations. For each sampled vector we recalculate the predicted equilibria thereby assigning bounds. We report 95% bootstrap bounds from 100 re-samplings.

The same approach applies to the 16 x 16 dinucleotide matrix with 240 parameters.

Comparing mutational matrices

We sought to test whether the predicted equilibria solutions were different between the matrices reflecting mutational profiles at 4-fold degenerate sites and all mutations at other sites (i.e. not 4 fold degenerate), as might be predicted were there contemporaneous selection against mutations that are non-synonymous. We partitioned all CDS mutations into those at 4-fold redundant sites (n=1151) and all others (n=5482). Using these two datasets we calculated observed equilibrium frequencies for each nucleotide (4* for 4-folds and n4* for non 4-folds, representing each as a vector of length four. We then determined the Euclidean distance between the two vectors. To test for significance, we compare the magnitude of this Euclidean distance to that expected by chance employing a non-parametric Monte Carlo simulation. To this end, we randomly extracted without replacement 1151 mutations from the full set of mutations so as to create a subsample of pseudo '4-folds'. The remaining 5482 mutations we then considered a sample of pseudo 'non 4-fold' mutations. For each randomization, we assembled the corresponding mutational matrix, solved for equilibria and calculated the Euclidean distance between the resulting vectors of predicted equilibrium for the four nucleotides. We repeated this procedure 10,000 times to generate a null distribution of Euclidean distances that controls for sample sizes differences. Significance was given as $P = n/m$, where n is the number of simulations in which the Euclidean distance is as great or greater than observed in the real data and m is the number of simulations (i.e. 10,000). To check for robustness, we considered an alternative distance metric, namely sum of modular differences (Euclidean distance considers square root of sum of squares of difference).

To consider each nucleotide individually, from the same Monte Carlo sampling we calculated the difference between predicted equilibria at sampled pseudo 'four-folds' and pseudo 'non four-folds' for the 10,000 repeats. This generates four distributions, one for each nucleotide. For each

nucleotide we calculate the mean (approximately zero) and standard deviation of these randomizations. The observed difference seen for each nucleotide between the equilibria predicted using mutations at four-fold sites (their predicted neutral equilibria) compared to that calculated using mutations at non four-fold site, may then be represented as a Z-score (Z=(observed – mean of simulations)/sd of simulations), $Z > |1.96|$ indicating significant deviation.

Homoplasy screen in SARS-CoV-2

Sites can appear as having independently occurring mutations for at least two reasons: the extra mutation may be a sequencing error or it may be a true homoplasy (i.e. the same mutation at the same site occurring more than once independently) (van Dorp, et al. 2020). Sequencing errors need to be removed. Knowing how to handle true homoplasies in the construction of a mutational matrix is not as conceptually simple.

At first sight one might suggest that, as independent mutations, each occurrence of the mutation should be considered. The key question, however, is whether the mutational profile at these sites is representative of activity at other sites. If it is not, then their over inclusion will bias the matrix towards the profile of homoplasic sites away from that of the rest of the genome, which could itself cause a false signal of non-equilibrium status (i.e. where mutationally predicted and observed nucleotide compositions – largely at non-homoplasic sites - disagree). A *priori* by virtue of the fact that they are homoplasic we might suppose that mutational activity at these sites is not reflective of the mutational profile elsewhere in the genome and it is the equilibrium properties of other sites that we are interested in. Equally these may well be sites that are more likely to be under selection (van Dorp, et al. 2020) and hence, again, not necessarily reflective of the mutational process. One could then opt to filter out mutations at homoplasic sites considering them possibly unrepresentative. However, we don't know they are unrepresentative and so their removal may be depleting the analysis of information. We also don't know how many of the non-mutated sites had had the property of being homoplasic prior to current sequencing. An alternative, the middle way, is to include them but count all occurrences at any given site as one event, thereby employing the mutations but preventing such sites from overly skewing the matrix and further reducing the impact of possible (missed) sequencing errors.

For analysis of SARS-CoV-2 we opt for the latter "middle way" approach but also check for resilience by removing such sites. Fortunately, as such sites are so rare (6 of 1151 4-fold degenerate sites), removal of these sites makes no important difference to calculation of GC equilibrium

content, nor to estimation of observed nucleotide content. We thus report the homoplasic-excluded results as minor asides.

Phylogenetic tree of 11,204 SARS-CoV-2 isolates was downloaded from the COVID-19 Genomics UK Consortium website (https://www.cogconsortium.uk/, version of 24-04-2020). Subsequently, the MSA and the resulting tree were used to identify recurrent mutations (homoplasies) using HomoplasyFinder (Crispell, et al. 2019). All ambiguous sites in the alignment were set to 'N'. Sites in the first and last 200 bp of the genome alignment were masked to account for the fact that a higher degree of spurious variants that can appear homoplasic tend to locate at the ends of the multiple sequence alignment.

HomoplasyFinder identified 408 putative homoplasies that were distributed over the SARS-CoV-2 genome. Homoplasies can occur as a result of convergent evolution, recombination, or due to artefacts such as specific combinations of sample preparation, sequencing technology, consensus calling approaches and sequencing errors. In order to remove spurious homoplasic sites, a particular worry of this dataset since a mix of technologies and methods have been employed by different contributing research groups, these were filtered using a set of parameters and thresholds defined in (van Dorp, et al. 2020) to obtain a set of high confidence homoplasies. Briefly, for each homoplasy, the proportion of isolates with the homoplasy where the nearest neighbouring isolate in the phylogeny also carried the homoplasy (pnn) was computed and all homoplasies with pnn < 0.1 were excluded. Furthermore, we also excluded homoplasies that were shared in less than 0.1% of the isolates (>11 isolates). We also required that no isolate had an ambiguous base near the homoplasies (±5 bp). These filters reduced the number of homoplasic sites to 67. The predicted equilibrium frequency of the four nucleotides at all the homoplasic sites (440 mutations), counting each class of mutation only once, is not different from that at non-homoplasic sites (Euclidean distance method: P=0.61). The filtered (accepted) homoplasies are also not significantly different from non-homoplasic mutations (Euclidean distance method: P=0.63) or from the rejected ones (Euclidean distance method: P=0.41). We conclude that the accepted set, with each mutation counted once, presents a defendable balance between inclusion and stringency.

Estimating dinucleotide enrichment

For the dinucleotide NpM (e.g. CpG, GpC etc), we define gene body enrichment (E(NM)) as:

$$E(NM) = p(NM)/[p(N) \times p(M)]$$

where p(NM) is the frequency of all dinucleotides within the gene that are NM and p(N) and p(M) are the frequencies of the mononucleotides within the same gene. We then consider site specific enrichment, i.e. sites 12, 23, or 31 defined by codon position, 31 being a third site and the codon first site of the following codon. Then at sites xy:

$E(NM_{xy}) = p(NM_{xy})/[p(N_x) \times p(M_y)]$

Where NM_{xy} is the relevant dinucleotide initiating with N at site x.

Gene expression

We employ expression data specified by Kim et al. (2020). We used the highest read count for each subgenomic RNAs in Supplemental Table 3 of Kim et al. (2020) and compared log2 normalised read counts to gene G+C content, G+C at third sites, and CpG enrichment. As the authors employed nanopore sequencing read count does not obviously require gene length normalization. Note that subgenomic RNA measures exclude ORF1a and ORF1b. ORF10 is excluded as no reads were identified. We used the shapiro.test function in R to test log2 transformed read counts for normality.

RNA stability

The minimum free energy of mRNA secondary structure was calculated for entire SARS-CoV-2 coding sequences (as defined in the "gene locations" section), using the hybrid-ss-min (UNAFold) program version 3.8 (Markham and Zuker 2008), with default settings (NA = RNA, t = 37). The folding energy of each sequence was then divided by the length of the corresponding sequence, to obtain the per nucleotide mRNA stability measure that was used in downstream calculations.

Data compilation of vertebrate viruses

Vertebrate virus sequences were retrieved from the Virosaurus database (Virosaurus databases 2020_4.1, Release April 2020, file: Virosaurus90v 2020_4.1)(Gleizes, et al. 2020) [accessed 07 May 2020]. In this database, complete sequences were clustered at 90% to remove redundancy. Since in this database, herpesviridae and poxviridae are split in genes rather than full genomes, complete sequences for these viruses were retrieved from NCBI RefSeq database (Pruitt, et al. 2005)). The same was also done for segmented viruses to allow calculation of sequence parameters per species. Genome classification was retrieved from ICTV Virus Metadata Repository: version May 1, 2020;

MSL35 (Walker, et al. 2019). Annotation for replication compartments was assigned according to ICTV (Walker, et al. 2019) and ViralZone (Hulo, et al. 2011) CpG and UpA enrichment were calculated as above. For virus sequences obtained from the Virosaurus database, the mean was derived to obtain one value per species. For segmented viruses, segments were first concatenated before calculating sequence parameters. Species information and sequence parameters can be found in Supplementary Table 2.

Genome sources

We acknowledge the sources of the genomes that we employed in Supplementary Table 3 (for SARS-CoV-2), Supplementary table 4 (for H1N1) and Supplementary table 5 (for Ebola).

RESULTS

SARS-COV-2 mutations are heavily GC→U biased

From the 14,599 genomes we can identify spontaneous mutations. From these we derive a mutational matrix and from this we solve for mutational equilibrium. From 1151 mutations at four-fold degenerate third sites we find a heavily GC→AU biased mutational profile (Table 1). From this we deduce that equilibrium GC (termed GC*) should be 17.13% (95% bootstrap estimates 17.09-17.52). The corresponding number is 17.10% on removing 6 homoplasies. Specifically, we find: U4*=65.67%; A4* = 17.20; C4* =13.09%; G4* = 4.04%. The striking bias towards U has been recently commented on and considered to be consistent with APOBEC editing (Di Giorgio, et al. 2020; Simmonds 2020).

	Derived allele				
Reference allele		**A**	**U**	**C**	**G**
	A	-	**0.04878** 0.02204	**0.01626** 0.01722	**0.12927** 0.10067
	U	**0.02454** 0.01753	-	**0.11528** 0.08912	**0.01296** 0.01296
	C	**0.05842** 0.03545	**0.54124** 0.40877	-	**0.01546** 0.00896
	G	**0.23913** 0.12389	**0.52174** 0.18060	**0.05072** 0.02111	-

Table 1 The 4 x 4 mutational matrix for 1151 mutations at four-fold synonymous sites (in bold) and from 5482 mutations observed anywhere in codons (not bold). Rates are defined as the number of observed changes per incidence of the nucleotide in the reference genome at four-fold third sites (bold) or in codons. Note that because of different normalizations, the two sets of numbers are not directly comparable in absolute terms.

Cognisant that there might be dinucleotide based mutation biases we extend the mononucleotide matrix to a 16 x 16 dinucleotide matrix with 240 parameter estimates derived across the coding sequences (Fig 1, Supplementary table 6). With 13209 dinucleotide switches this represents an average of 55.04 mutations per parameter estimate which is liable to be noisy and potentially weakly influenced by selection on non-synonymous mutations. With this we determine equilibrium content for all dinucleotides and in turn all nucleotides (A*=17.66%, C*=11.65%, U *=62.42%, G*=8.27%). We thus estimate from this GC* of 19.92% (95% bootstraps 19.87-20.05%) more or less in line with mononucleotide calculations.

Evidence for selection acting to counter a large mutation bias towards U

If selection favours reduced G+C content owing to selection for reduced CpG content, we expect that the observed GC3 should be lower than that predicted under neutrality (17.1%). We find the opposite to be true, observed GC3 being 28% (GC3 at 4-fold sites = 20.2%). All numbers are beyond 95% bootstrap bounds of the predicted equilibrium frequency derived from analysis of mononucleotide profiles at 4-fold degenerate sites (bounds: 17.09-17.52). More specifically, at four-fold synonymous sites, observed U4 (50.8%) is less than predicted under neutral equilibrium U4* (65.7%) while all other bases are higher than expected (A4 = 28.95%, A4* = 17.20%; C4 = 13.70%, C4* =13.09%; G4 = 6.50%, G4* = 4.04%.). A parsimonious explanation is that the sizeable mutation bias towards U generates deleterious mutations, non-optimal even at synonymous sites, and selection therefore favours reduced U content.

GC of coding sequence is even more removed from the neutral equilibrium at 38%. This deviation suggests selection in favour of non-synonymous mutations that increase G+C content. Examination of non-equilibrium status by dinucleotide content supports this. It shows one striking effect, namely that UU's predicted equilibrium frequency greatly exceeds what is observed (Figure 2). More generally, U content whether derived from mononucleotides at 4-fold third sites (predicted 65.7%) or mononucleotides across the genes (predicted 60.3%) or from dinucleotides (62.4%) is greatly in excess of U content, this being 32% for the complete viral sequence. The

mutational matrix, whether through mono or dinucleotide analysis, predicts a great enrichment of U which we infer is being opposed by selection at third sites and in gene bodies (unweighted gene body means: U1% = 25.7%, U2%=36.3%, U3%=41%). We notice that CpG content is above that expected under neutrality (Figure 2). However, this we suggest is not so much evidence against selection towards high CpG so much as selection against UU, which by necessity increases the observed relative frequency of CpG and most other dinucleotides as frequencies must sum to one.

Evidence for contemporaneous selection against U at non-four-fold redundant sites

Possibly consistent with a role for selection, using 5482 mutations that occur anywhere in the coding sequence (Table 1), we observe that the G→U flux at 4-fold degenerate sites is much greater than that observed throughout the sequence. The same is true to a lesser extent for the C→U and A→U fluxes. Assuming the flux rate at 4-fold degenerate sites is more indicative of the true mutational flux, this is consistent with non-synonymous U mutations being under strong enough selection to be eliminated prior to sequencing. The predicted bias from this matrix is thus slightly more GC rich than that determined from 4-fold redundant sites (GC* at 21.37%: 95% bootstrap estimates 21.39-21.60, 21.44% also after excluding homoplasies).

The lower occurrence of mutations generating U at non-4-fold sites would be consistent with contemporary selection on non-4-fold sites opposing mutations towards U, consistent also with the different between U4*, U4 and U content overall. To ask whether the difference between the two equilibria solutions is significantly different, we developed a non-parametric Monte Carlo simulation (Methods). We find that the Euclidean distances from the random sampling are the same as, or greater than, the Euclidean distance between four-folds and non-four fold sites in just 323/10,000 cases (hence $P = 0.0323$) (repeating using an alternative distance metric, sum of modular difference between equilibria, make no meaningful difference, $P=0.0454$). To clarify that it was selection against U, we considered each nucleotide individually (see Methods). Such analysis indeed provides evidence for significant counter selection of U at non-4-fold sites (Un4*=60.8, U4 *=64.6%, Z = -1.98). Commensurably, predicted G equilibrium content derived from mutations at non-4-fold sites is higher than that derived from mutations at 4-fold degenerate sites (Z = 5.34), while A and C content are less affected (Z for A = 0.26, Z for C = -0.56). Thus, not only do we detect deviation away from the predicted neutral equilibrium (at 4-fold sites, third sites generally and through the gene body), we also can detect a signal consistent with selection on SARS-CoV-2 that skews the mutational matrix prior to sequencing.

Significant heterogeneity in the degree of CpG avoidance between genes

While selection against U or UU provides a viable model for GC3>GC3*, might there be other explanations that would be consistent with selection against CpG, to avoid ZAP, but in favour of G+C? One possibility is that we may be witnessing between-gene heterogeneity (Digard, et al. 2020). Imagine that some genes are indeed under selection for low CpG and hence for low GC3, but others are not under selection for low CpG and thus are more free to have selection favouring higher GC3 (for unspecified reasons, but possibly to enable efficient expression (Mordstein, et al. 2020)). When then considered *en masse* we see both selection for CpG and more raised GC3. Recent reports suggest that not all genes are equally subject to selection for low CpG to avoid ZAP, with only "immediate early" genes under such selection (Lin, et al. 2020).

Were this the explanation, or part thereof, we would predict that CpG enrichment would be heterogeneous between genes (see also Digard, et al. 2020) and that those with relatively high CpG enrichment will also be those genes contributing to raised GC3 (i.e. a positive correlation between CpG enrichment and GC3). Note that while CpG counts are likely to be necessarily higher as GC3 goes up, CpG enrichment is normalised to underlying GC content and so CpG enrichment and high GC3 are not logically coupled (e.g. if at the limit 50% of residues are C and 50% G, so long as CpG usage is random, CpG = 0.5*0.5, CpG enrichment will not be seen).

To assay this, we calculated CpG enrichment at codon sites 12, 23 and 31, these providing three measures of CpG enrichment for each gene. We can then perform a Kruskal-Wallis test for heterogeneity. Even with such limited data, we find that the three measures for the same gene are more similar than expected by chance (KW, P=0.019, df=11: mean E(CG) =0.61 +/- 0.4 sd; Fig 3a). This implies that at all sites CpG is avoided or preferred to the same degree within any given gene. We see however only marginal evidence that genes released from CpG constraint are those with higher GC3 (CpG enrichment v GC3, rho =0.41, P=0.19, Spearman's test, Fig 3b). Thus, while there is evidence for differential CpG usage between genes, we don't find that this predicts GC3, although trends are in the expected direction and the tests underpowered.

More generally we can ask whether gene body G+C content behaves the same as gene body CpG content with each gene having its own characteristic profile. We assay this by considering GC1, GC2 and GC3 in a manner as above. We find no evidence that genes are more similar in these three measures than expected by chance (KW P=0.49, df=11: Fig 4). Similarly, we see no correlation between GC3 and GC12 although the trend is positive (rho = 0.15, P=0.63,

Spearman's rank) (see also Dilucca et al (2020)). However, we do observe some regularities. First, GC3 is consistently lower than GC12 (Wilcoxon signed-rank test, P=0.007), the mean GC3 being 28%, while that of GC12 is 40%, consistent with selection on amino acid content.

The most striking feature of third site nucleotide usage is that all genes have a preponderance of U (Figure 5). As noted above, this we can attribute only in some part to mutation as the predicted levels while in the rank order as observed (U>A>C>G) are highly deviant from null. Specifically, the predicted numbers are 0.66>0.17>0.13>0.04 while the observed are 0.44>0.28>0.16>0.13. Approximately the same predicted equilibrium values are seen employing all mutations (0.60>0.18>0.13>0.08). Selection against U seems strong, despite this being the most common nucleotide, as it is heavily reduced from its predicted equilibrium content.

Genes avoiding CpG also avoid UpA

Prior analysis suggests that viruses lacking CpG also tend not to have UpA and that engineering increased CpG and UpA attenuates viruses, possibly because both are under-represented in human transcripts (Simmonds, et al. 2013). We also observe that UpA enrichment and CpG enrichment tend to positively correlate across viruses (N=1290, rho=0.165, P=2.68 x $10^{-9;}$ data in Supplementary Table 2). To understanding whether increasing CpG and UpA might be a useful attenuation strategy, we ask whether UpA is also avoided in genes of SARS-CoV-2 and whether it is avoided in the same genes that avoid CpG. We consider not just the CpG enrichment predicting UpA enrichment but also, to control for mononucleotide effects, the two other symmetric nucleotide pairings (ApU and GpC).

	UpAe	**ApUe**	**CpGe**	**GpCe**
UpAe	-	0.20	**0.76****	-0.18
ApUe		-	-0.15	-0.16
CpGe			-	0.007
GpCe				-

Table 2 Between-gene correlations in dinucleotide enrichment scores (Pearson product moment correlation r values). Significant correlations in bold: ** = P <0.005.

On the average UpA is, like CpG, avoided although not to the same extent as CpG (mean UpA enrichment =0.83 +/- 0.2 sd) (Figure 6b). UpA also shows between gene heterogeneity (KW test P=0.04). We find that exclusively for CpG enrichment and UpA enrichment do we see a

correlation between genes (Table 2, Fig 6a). ApU is also avoided (mean enrichment = 0.83 +/- 0.14 sd), but there is no evidence for within gene homogeneity (KT test P=0.14) (Fig 7). By contrast, there is no evidence for GpC avoidance: mean GpC enrichment =1.13, +/- 0.34 sd, Fig 7) and genes do not show gene-specific GpC enrichment (KW, P=0.11, comparing GpC enrichment at sites 12, 23 and 31). We conclude that if CpG enrichment is a viable strategy to attenuate a gene, increasing UpA may also (although for reasons unknown).

Evidence for U content predicting expression level.

The results above are consistent with a model in which CpG content is under selection in some genes to be reduced, while GC3 content is above the level expected under neutrality, in no small part because the U mutation bias is so extreme that equilibrium U content (especially UU content) would render the virus much less fit. There are several possible mechanistic explanations for the GC3 > GC3* effect. With our recent evidence that intronless low GC genes are barely expressed in human cell lines (Mordstein, et al. 2020), selection for raised GC3 (reduced U3) to enable more effective gene expression is a strong contender. In this context, while we do not see a GC3 expression correlation (r=0.09, P=0.82), we do observe a GC expression correlation (r=0.79, p=0.01 and figure 8). Breaking this down by nucleotide we see that this is owing to a negative correlation with U content and a positive correlation with both C and G content (A freq: r= 0.33, P=0.83; C freq: r=0.64, P=0.06; G freq r=0.81, P=0.009; U freq r=-0.88, P=0.0017). Why this is will require considerable experimental manipulation of sequences to understand but we note a correlation between expression level and predicted per nucleotide stability (Pearson's r= -0.86, p = 0.0027, df = 7). It is notable that we observe such an effect with such an underpowered test.

A more broad-brush approach is to consider viral sequences more generally (Supplementary Table 2). As part of the mechanism by which GC enrichment boosts expression is thought to be intranuclear (e.g nuclear export) (Mordstein, et al. 2020), if selection is operating on gene expression of viruses, we might predict that nuclear viruses might have a higher GC content than cytoplasmic viruses. Using mean GC of all viruses within a taxonomic grouping we observe this to be the case (Mann Whitney U test $P<2.2 \times 10^{-16}$, Fig 9). CpG enrichment and UpA enrichment is similarly lower in cytoplasmic viruses (Fig 9). This is a very arms-length result and requires due caution in its interpretation (it could just as well be evidence of different mutational biases). Nonetheless, within the context of our prior result we suggest that this merits further scrutiny. There is some evidence that if selection might favour reduced CpG content it might also favour reduced UpA content, as, within both groups, those viruses with low CpG enrichment also tend

to have low UpA enrichment, but the effect is weak (spearman's test, cytoplasmic viruses: rho=0.096, p=0.016; nuclear viruses: rho=0.084, p=0.031).

Designing the optimally attenuated SARS-CoV-2

With the above evidence for selection for G+C at third sites and for heterogeneity between genes in enrichment of CpG and UpA, we suggest that simply increasing CpG by manipulation of synonymous sites need not be the optimal strategy. It may enable recognition by ZAP, but may also favour increased fitness by increasing G+C/reducing U.

As not all genes are under selection for reduced CpG/UpA, reducing their G+C content by increasing U content seems a relatively safe and robust strategy. We thus suggest to classify genes according to the CpG enrichment (>1 or <1). For those in the first category, likely not affected by ZAP (E and ORF10, (see also Digard, et al. 2020)) we suggest decreasing their synonymous G+C by increasing where possible U content and forcing them closer to their mutational equilibrium. For those with especially low CpG enrichment and most likely strong targets of ZAP (ORF1a, ORF1b, ORF6, ORF7b and S) we suggest, raising their CpG, even at the cost of increased G+C. Where possible UpA should also be increased. For the remainder we suggest increasing CpG content while holding GC3 content static or decreasing if possible. However, with the possibility of synonymous sites also being parts of key motifs, e.g. for RNA binding proteins (Savisaar and Hurst 2017), a simplistic strategy, even if gene-tailored, may have deleterious undesirable side consequences. Unlike alternative strategies that permute existing codons (Jorge, et al. 2015) the proposed strategy (Supplementary 7) enables deviations in overall nucleotide content. Recognizing, however, that the extreme nucleotide content can cause gene inactivation, as with prior strategies (Jorge, et al. 2015), we propose a stochastic methodology to derive a large number of variants modified in the desired direction that could be experimentally tested (for algorithm see Supplement 7; for variants for each gene see Supplement 8). We suggest variants for each gene recognizing the most effective attenuated construct may be a mosaic of wild-type and attenuated genes (Coleman, et al. 2008).

GC3*>GC3 is not a general property of viruses

We observed that GC content at third sites was both higher than expected given selection against CpG and higher than expected given the underlying mutational profile. Is a deviation from mutational equilibrium a general property of human viruses? Were this so, this too could have implications for engineering of attenuated forms. To address this, we consider other viruses with rich sequencing from epidemics.

For H1N1 using the same mode of analysis we observe both a less extreme GC->AT mutation bias (Table 3) and an observed GC3 content very close to that predicted. From analysis of 3rd sites the predicted value is GC3*=41.8% (bootstrap 95% intervals 41.45-42.04), from all sites the prediction is GC*=42.8% (bootstrap 95% 42.56-42.96). The observed GC3 is 41.8%, within the bounds of the prediction based on 3rd site mutations. For Ebola (Table 4) we find observed GC3 is all but identical to predicted (observed GC3=46.4%, expected=46.7%). We conclude that analysis of SARS-CoV-2 and its non-mutational equilibrium status at synonymous sites does not necessarily hold lessons for other viruses. In contrast to others (Kames, et al. 2020), we suggest caution in generalizing vaccine strategies.

	Derived allele			
Reference allele	**A**	**U**	**C**	**G**
A	-	**0.0871** 0.04597	**0.065** 0.0451	**0.4291** 0.25542
U	**0.0803** 0.05143	-	**0.4945** 0.24889	**0.0529** 0.03429
C	**0.1691** 0.11426	**0.5699** 0.30675	-	**0.0251** 0.02607
G	**0.6089** 0.32052	**0.0948** 0.05027	**0.0323** 0.0207	-

Table 3 The 4 x 4 mutational matrix for 1522 mutations at synonymous sites (in bold) and from 2571 mutations observed anywhere in codons (not bold) for H1N1. Rates are defined as the number of observed changes per incidence of the nucleotide in the reference genome at 3rd sites (bold) or in codons.

	Derived allele			
		A	**U**	**C**	**G**

Reference allele	**A**	**-**	**0.0739** 0.05077	**0.0964** 0.06722	**0.2123** 0.14803
	U	**0.0594** 0.05152	**-**	**0.2145** 0.13429	**0.0536** 0.04786
	C	**0.0845** 0.08086	**0.2639** 0.14868	**-**	**0.0394** 0.04845
	G	**0.2639** 0.16051	**0.0751** 0.05139	**0.0694** 0.05139	**-**

Table 4 The 4 x 4 mutational matrix for 1682 mutations at synonymous sites (in bold) and from 3523 mutations observed anywhere in codons (not bold) for Ebola. Rates are defined as the number of observed changes per incidence of the nucleotide in the reference genome at 3rd sites (bold) or in codons.

DISCUSSION

Mutation bias across all taxa is typically GC->AT biased (Hershberg and Petrov 2010; Hildebrand, et al. 2010; Liu, et al. 2018) and neutral predicted equilibrium frequencies below GC of 20% (as observed here) are not without precedent (see e.g. (Long, et al. 2018)). Broadly the U enrichment at third sites within the genome is then compatible with a large role for mutation bias, possibly mediated by members of the APOBEC gene family (Di Giorgio, et al. 2020; Simmonds 2020), known mutators of viruses (Lee, et al. 2008) with C→U and UC→UU preferences (Chen and MacCarthy 2017). However, we have shown that nucleotide usage, while skewed in the direction imposed by mutation bias, is nonetheless deviant from it. The difference between observed and expected U3 and UU (Fig 2) proportions are noteworthy. At four fold degenerate sites while C and G usage are close to equilibrium, A is far above and U is far below (U4=50.8%, U4*=65.67%; A4 = 28.95%, A4* = 17.20; C4 = 13.70%, C4* =13.09; G4 = 6.50%, G4* = 4.04%.). We propose that a parsimonious explanation is that the sizeable mutation bias towards U generates deleterious mutations, even at synonymous sites, and selection therefore favours reduced U content. However, increasing C or G potentially comes at a cost of increased CpG, so the base most in excess of its equilibrium is A. As a consequence, while CpG avoidance is real in some genes, GC3 is a little higher than predicted from the underlying mutational profile. This thus presents an unusual case in which the most common synonymous codons (those ending in U) are not the selectively advantageous ones.

We haven't directly addressed the problem of the causes of any such selection on synonymous mutations. Given G+C preference in human coding genes to enable effective expression (Kudla, et al. 2006; Mordstein, et al. 2020), the negative correlation between U usage and expression is broadly consistent with evidence for preferential degradation of transcripts with non-optimal codon usage (Radhakrishnan, et al. 2016; Buschauer, et al. 2020). Potentially in tandem to such possible effects high U content may trigger immunogenicity of nucleic acids via TLR-7 (Diebold, et al. 2004). Whether a virus with a few more U residues is importantly more immunogenic is, however, uncertain. Alternatively, effects may be mediated by changes in mRNA secondary structure (Mauger, et al. 2019). We indeed observe a correlation between expression level and predicted per nucleotide stability. Given this, it could be speculated that RNA stability may explain the thermal intolerance of the virus (Demongeot, et al. 2020), although many other mechanisms are imaginable.

On a related note, aside from the possible influence of APOBEC generating the excess of UU mutations, we have not considered the causes of the very different rates for each class of mutation. Indeed, for the high G→U rate we know of no editing process that has this profile (see Supplementary Table 1). However, we speculate that this might be owing to oxidation of guanosine that can lead to a G to U transversion. This process may be accelerated by NO via its oxidate species $ONOO^-$ (Yermilov, et al. 1995; Juedes and Wogan 1996), the former being produced primarily by inducible NO Synthase (iNOS). This enzyme is upregulated in many cells including inflammatory phagocytic cells including macrophages, mediated by proinflammatory cytokines including IFNγ (Zhuang, et al. 1998). This has known effects on viral mutation (for review see Akaike and Maeda 2000). It may also be informative to consider the relationship between RNA secondary structure and these mutation biases (Krishnan, et al. 2004), although overall proportion of variable 4-fold redundant sites doesn't covary with stability (Pearson's $r = -0.06$, $p = 0.88$, $df = 7$).

There are, however, at least four problems with our mode of analysis. First, a theoretical alternative explanation for the difference between predicted and observed values is that the virus was at neutral mutational equilibrium in its prior host (cf. H1N1, Ebola), but since the transfer to humans the mutational profile has altered. Were this so we may just have identified a lag in viral evolution from one neutral equilibrium to another. In this context deviation from equilibrium has little if anything to say about either selection or optimal vaccine design. While evidence for GC→AT biased mutation in related viruses (Simmonds 2020) renders this less parsimonious an explanation,

direct examination of mutational profiles of the virus in its ancestral host (whatever that may be) would be valuable. The evidence for subtly but significantly different mutational matrices dependent on the class of site employed provides more direct evidence for contemporary selection on U content throughout gene bodies that cannot be accounted for by a temporal shift in mutational profile.

Second, assuming no change to the mutational matrix, *sensu stricto*, we have observed a force that would cause a fixation bias (Lercher, et al. 2002). Evidence for such a force need not necessarily indicate the direction of selection, as selection bias is only one class of fixation bias. In biased gene conversion, for example, the mismatch repair machinery recognises, during double strand break repair, heteroduplex GC:AT mismatches and corrects these in favour of GC residues (Brown and Jiricny 1988). This causes a meiotic drive like process in which deleterious mutations can be driven to higher frequencies (for further consideration see Hurst 2019). Given that single strand RNA cytoplasmic viruses, such as SARS-CoV-2, are unlikely to be exposed to the nuclear mismatch repair machinery or need double strand break repair, biased gene conversion is unlikely to explain GC3>GC3*, U3<U3* *etc.* We cannot with our data, however, rule out unknown mechanisms causing similar non-selective fixation biases. It is then valuable to provide more direct evidence for an advantageous effect of reduced U3/increased GC3, as suggested by our preliminary analysis on expression level. Experimental manipulation of GC3 content (cf. Kudla, et al. 2006; Mordstein, et al. 2020) is a high priority.

Third, we have presumed that the mutational spectrum observed at 4-fold degenerate sites is a good reflection of the true mutational profile. Often when applying methodology like this we presume that the temporal proximity between occurrence and observation of mutations is so small that there has been no time for selection to filter in a manner that distorts the mutational matrix. Nonetheless, we found that although slight, there is a difference between the mutational profile observed at CDS sites that are not 4-fold degenerate and those that are. While this difference is so slight it cannot explain why U is so deviant from equilibrium levels, and doesn't question our overall findings, we do nonetheless presume that the 4-fold site matrix itself is unbiased. For strains sequenced hours to days apart to be biased at 4-fold degenerate sites would require strong and biased selection at 4-fold redundant sites. While not obviously plausible we have no means to disprove this (and strong selection, albeit associated with splicing, has been identified at synonymous sites in human genes (Savisaar and Hurst 2018)). Nonetheless, any such bias would also force the mutational matrix observed to predict a nucleotide content more closely resembling

the observed nucleotide content, rendering the test conservative. Derivation of the mutational profile by *in vivo* analysis (cf. Denison, et al. 2011) could enable more direct tests of our findings. Analysis of SARS-CoV (responsible for the 2002 SARS outbreak) with exonuclease activity (which we presume to mimic SARS-CoV-2, it having a nsp14 homolog of the SARS-CoV exonuclease (Pachetti, et al. 2020)) reports a massive AU→GC bias with 8 of 11 reported mutations being in this direction and only 2 GC→AU (Smith, et al. 2013). This implies either a radically different mutation bias in SARS-CoV than in SARS-CoV-2 or great sensitivity of results to experimental conditions, such as cell lines employed and APOBEC activity. We note that the SARS-CoV analysis employed Vero cells in which the interferon response is disabled (Smith, et al. 2013), thus likely to have neither ZAP (see e.g. MacDonald, et al. 2007) nor APOBEC activity (see e.g. Peng, et al. 2006).

Fourth, we have presumed that, after filtering (see Methods), all sequences are error free. While sequencing errors cannot explain a bias as strong as the difference between excess and expected UU or U3, nor can they obviously explain the evidence for contemporary selection against U, it may possibly explain the small difference between predicted and observed nucleotide content at 4-fold sites for G and C (the deviations of A and U from predicted equilibria are relatively large). One suggested means to avoid this is to only employ mutations that have been sequenced more than once (Hildebrand, et al. 2010). However, this has been shown to introduce its own bias (Charneski, et al. 2011). Using high quality sequence, it was shown that using mutations that appear once and those that appear twice or more makes a significant difference to the matrix and estimates of equilibria (Charneski, et al. 2011). The cause of this is likely to be a selection filter: mutations that persist longer to be sequenced twice or more will be skewed towards milder effect mutations. This accords with our observation of a slight difference between matrices that restrict just to 4-fold degenerate sites and those that do not. The ideal then is to filter not by regularity of appearance but by sequencing quality (hence our decisions on which sequences to employ: see Methods). Nonetheless, to err on the side of caution we considered mutations at 4-fold degenerate third sites that appear more than once (i.e. excluding singletons) and found that GC* is now even lower than previously predicted (GC*=10.3%, 95% bounds 10.19 -10.61). Thus, we are confident that we can exclude sequencing error as an explanation for observed GC3 > GC3* and U4<U4* (singleton excluded prediction of U4*=65.7%). Nonetheless, owing to observation bias and low sample size we caution against over-interpretation of this result. Given possible biases owing to sequencing platform, we also ask about the expected equilibrium content for illumina and nanopore

sequencing separately. We find that the predicted equilibrium vectors for 4-fold degenerate sites are no different from each other (P=0.62).

Assuming we have identified the direction of selection (against U, against CpG in some genes) this can inform vaccine design. Unusually, even though U is the most common nucleotide at third sites (by a considerable margin), we propose increasing this even more thereby forcing the viruses against the direction of purifying selection. We predict that raising CpG in the genes that are CpG deficient would be a viable strategy even at a cost of raising GC3/lowering U. By contrast for those few genes with E(CpG)>1 (i.e. gene E, ORF10, see also Digard et al. (2020)) CpG manipulation increasing GC3 would be a dangerous strategy, potentially achieving little more than an increase in expression. Increasing their U content would appear to be the anti-selection direction. We note however that ORF10's function, if any, remains unclear there being no evidence of transcripts from it, despite it looking like a well formed ORF (starts ATG stops TAG, multiple of 3 long). Its GC3 content is also far from neutral equilibrium (GC3=36%). In this context gene E may be a good one to alter synonymous site usage as it appears not to be under selection for CpG or UpA avoidance.

Genes ORF1a, ORF1b, ORF6, ORF7b and S are good candidates for the raising of CpG content. Gene N is noteworthy in being very highly expressed, long (1260 bp), GC rich (GC3=38%) and with moderate CpG enrichment (E(CpG)=0.56). Given these characteristics it should be possible to increase CpG by manipulating some third sites (those with C at codon position 2 or G at codon position +1) while reducing GC and increasing U content at other sites. For smaller genes there is less leeway. In this context S, ORF1a and ORF1b are also very strong candidates being long, with moderate GC3 and low CpG enrichment. A more detailed description of the algorithm for attenuation alongside attenuated variants can be found in Supplementary Tables 7 and 8. While the particular strategy for attenuation reflects the particulars of selection operating on SARS-CoV-2, the more general notion of evolutionarily-informed vaccine design, with attenuation achieved by synthesising variants rich in the compositional features opposed by selection, is worthy of experimental scrutiny.

Acknowledgements

This work was supported by the Wellcome Trust (fellowship 207507 to G.K.) and the European Research Council (advanced grant ERC-2014-ADG 669207 to L.D.H.)

References

Abascal F, Zardoya R, Telford MJ. 2010. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic Acids Research 38:W7-13.

Akaike T, Maeda H. 2000. Nitric oxide and virus infection. Immunology 101:300-308.

Amanat F, Krammer F. 2020. SARS-CoV-2 Vaccines: Status Report. Immunity 52:583-589.

Bentele K, Saffert P, Rauscher R, Ignatova Z, Bluthgen N. 2013. Efficient translation initiation dictates codon usage at gene start. Molecular Systems Biology 9:675.

Brown TC, Jiricny J. 1988. Different base base mispairs are corrected with different efficiencies and specificities in monkey kidney-cells. Cell 54:705-711.

Buschauer R, Matsuo Y, Sugiyama T, Chen YH, Alhusaini N, Sweet T, Ikeuchi K, Cheng J, Matsuki Y, Nobuta R, et al. 2020. The Ccr4-Not complex monitors the translating ribosome for codon optimality. Science 368:eaay6912.

Cai Y, Ye C, Cheng B, Nogales A, Iwasaki M, Yu S, Cooper K, Liu DX, Hart R, Adams R, et al. 2020. A Lassa Fever Live-Attenuated Vaccine Based on Codon Deoptimization of the Viral Glycoprotein Gene. Mbio 11:e00039-00020.

Carneiro J, Pereira F. 2016. EbolaID: An Online Database of Informative Genomic Regions for Ebola Identification and Treatment. PLoS Negl Trop Dis 10:e0004757.

Charneski CA, Honti F, Bryant JM, Hurst LD, Feil EJ. 2011. Atypical at skew in firmicute genomes results from selection and not from mutation. PLoS Genetics 7:e1002283.

Chen J, MacCarthy T. 2017. The preferred nucleotide contexts of the AID/APOBEC cytidine deaminases have differential effects when mutating retrotransposon and virus sequences compared to host genes. PLoS Computational Biology 13: e1005471.

Coleman JR, Papamichail D, Skiena S, Futcher B, Wimmer E, Mueller S. 2008. Virus attenuation by genome-scale changes in codon pair bias. Science 320:1784-1787.

Crispell J, Balaz D, Gordon SV. 2019. HomoplasyFinder: a simple tool to identify homoplasies on a phylogeny. Microb Genom 5:mgen.0.000245.

Demongeot J, Flet-Berliac Y, Seligmann H. 2020. Temperature Decreases Spread Parameters of the New Covid-19 Case Dynamics. Biology 9:94.

Denison MR, Graham RL, Donaldson EF, Eckerle LD, Baric RS. 2011. Coronaviruses: an RNA proofreading machine regulates replication fidelity and diversity. RNA Biol 8:270-279.

Di Giorgio S, Martignano F, Torcia MG, Mattiuz G, Conticello SG. 2020. Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. Science Advances 6:eabb5813.

Diebold SS, Kaisho T, Hemmi H, Akira S, Reis e Sousa C. 2004. Innate antiviral responses by means of TLR7-mediated recognition of single-stranded RNA. Science 303:1529-1531.

Digard P, Lee H-M, Sharp C, Grey F, Gaunt ER. 2020. Intra-genome variability in the dinucleotide composition of SARS-CoV-2. bioRxiv:2020.2005.2008.083816.

Dilucca M, Forcelloni S, Georgakilas AG, Giansanti A, Pavlopoulou A. 2020. Codon Usage and Phenotypic Divergences of SARS-CoV-2 Genes. Viruses 12:498.

Elson D, Chargaff E. 1952. On the desoxyribonucleic acid content of sea urchin gametes. Experientia 8:143-145.

Fath S, Bauer AP, Liss M, Spriestersbach A, Maertens B, Hahn P, Ludwig C, Schafer F, Graf M, Wagner R. 2011. Multiparameter RNA and codon optimization: a standardized tool to assess and enhance autologous mammalian gene expression. Plos One 6:e17596.

Ficarelli M, Antzin-Anduetza I, Hugh-White R, Firth AE, Sertkaya H, Wilson H, Neil SJD, Schulz R, Swanson CM. 2020. CpG Dinucleotides Inhibit HIV-1 Replication through Zinc Finger Antiviral Protein (ZAP)-Dependent and -Independent Mechanisms. Journal of Virology 94:e01337-01319.

Gaunt E, Wise HM, Zhang H, Lee LN, Atkinson NJ, Nicol MQ, Highton AJ, Klenerman P, Beard PM, Dutia BM, et al. 2016. Elevation of CpG frequencies in influenza A genome attenuates pathogenicity but enhances host response to infection. Elife 5:e12735.

Virosaurus [Internet]. 2020. Available from: https://viralzone.expasy.org/8676

Gustafsson C, Govindarajan S, Minshull J. 2004. Codon bias and heterologous protein expression. Trends in Biotechnology 22:346-353.

Hershberg R, Petrov DA. 2010. Evidence That Mutation Is Universally Biased towards AT in Bacteria. PLoS Genetics 6:e1001115.

Hildebrand F, Meyer A, Eyre-Walker A. 2010. Evidence of selection upon genomic GC-content in bacteria. PLoS Genetics 6:e1001107.

Phylodynamic analysis of SARS-CoV-2 | Update 2020-03-06 [Internet]. 2020. Available from: https://virological.org/t/phylodynamic-analysis-of-sars-cov-2-update-2020-03-06/420

Hulo C, de Castro E, Masson P, Bougueleret L, Bairoch A, Xenarios I, Le Mercier P. 2011. ViralZone: a knowledge resource to understand virus diversity. Nucleic Acids Research 39:D576-582.

Hurst LD. 2019. A century of bias in genetics and evolution. Heredity (Edinb) 123:33-43.

Jorge DM, Mills RE, Lauring AS. 2015. CodonShuffle: a tool for generating and analyzing synonymously mutated sequences. Virus Evol 1:vev012.

Juedes MJ, Wogan GN. 1996. Peroxynitrite-induced mutation spectra of pSP189 following replication in bacteria and in human cells. Mutation Research-Fundamental and Molecular Mechanisms of Mutagenesis 349:51-61.

Kames J, Holcomb DD, Kimchi O, DiCuccio M, Hamasaki-Katagiri N, Wang T, Komar AA, Alexaki A, Kimchi-Sarfaty C. 2020. Sequence analysis of SARS-CoV-2 genome reveals features important for vaccine design. bioRxiv:2020.2003.2030.016832.

Karlin S, Doerfler W, Cardon LR. 1994. Why Is Cpg Suppressed in the Genomes of Virtually All Small Eukaryotic Viruses but Not in Those of Large Eukaryotic Viruses. Journal of Virology 68:2889-2897.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biology and Evolution 30:772-780.

Kerns JA, Emerman M, Malik HS. 2008. Positive selection and increased antiviral activity associated with the PARP-containing isoform of human zinc-finger antiviral protein. PLoS Genetics 4:ARTN e21.

Khan KH. 2013. DNA vaccines: roles against diseases. Germs 3:26-35.

Kim D, Lee JY, Yang JS, Kim JW, Kim VN, Chang H. 2020. The Architecture of SARS-CoV-2 Transcriptome. Cell 181:914-921 e910.

Krishnan NM, Seligmann H, Raina SZ, Pollock DD. 2004. Detecting gradients of asymmetry in site-specific substitutions in mitochondrial genomes. DNA and Cell Biology 23:707-714.

Kudla G, Lipinski L, Caffin F, Helwak A, Zylicz M. 2006. High guanine and cytosine content increases mRNA levels in mammalian cells. PLoS Biology 4:933-942.

Kudla G, Murray AW, Tollervey D, Plotkin JB. 2009. Coding-Sequence Determinants of Gene Expression in Escherichia coli. Science 324:255-258.

Lee YN, Malim MH, Bieniasz PD. 2008. Hypermutation of an ancient human retrovirus by APOBEC3G. Journal of Virology 82:8762-8770.

Lercher MJ, Smith NG, Eyre-Walker A, Hurst LD. 2002. The evolution of isochores. evidence from SNP frequency distributions. Genetics 162:1805-1810.

Lin Y-T, Chiweshe S, McCormick D, Raper A, Wickenhagen A, DeFillipis V, Gaunt E, Simmonds P, Wilson SJ, Grey F. 2020. Human cytomegalovirus evades ZAP detection by suppressing CpG dinucleotides in the major immediate early genes. bioRxiv:2020.2001.2007.897132.

Liu HX, Huang J, Sun XG, Li J, Hu YW, Yu LY, Liti GN, Tian DC, Hurst LD, Yang SH. 2018. Tetrad analysis in plants and fungi finds large differences in gene conversion rates but no GC bias. Nature Ecology & Evolution 2:164-173.

Long H, Sung W, Kucukyildirim S, Williams E, Miller SF, Guo W, Patterson C, Gregory C, Strauss C, Stone C, et al. 2018. Evolutionary determinants of genome-wide nucleotide composition. Nat Ecol Evol 2:237-240.

MacDonald MR, Machlin ES, Albin OR, Levy DE. 2007. The Zinc Finger Antiviral Protein Acts Synergistically with an Interferon-Induced Factor for Maximal Activity against Alphaviruses. Journal of Virology 81:13509-13518.

Manokaran G, Sujatmoko, McPherson KG, Simmons CP. 2019. Attenuation of a dengue virus replicon by codon deoptimization of nonstructural genes. Vaccine 37:2857-2863.

Markham NR, Zuker M. 2008. UNAFold: software for nucleic acid folding and hybridization. Methods Mol Biol 453:3-31.

Mauger DM, Cabral BJ, Presnyak V, Su SV, Reid DW, Goodman B, Link K, Khatwani N, Reynders J, Moore MJ, et al. 2019. mRNA structure regulates protein expression through changes in functional half-life. Proc Natl Acad Sci U S A 116:24075-24083.

Mordstein C, Savisaar R, Young RS, Bazile J, Talmane L, Luft J, Liss M, Taylor MS, Hurst LD, Kudla G. 2020. Codon Usage and Splicing Jointly Influence mRNA Localization. Cell Syst 10:351-362 e358.

Mueller S, Coleman JR, Papamichail D, Ward CB, Nimnual A, Futcher B, Skiena S, Wimmer E. 2010. Live attenuated influenza virus vaccines by computer-aided rational design. Nature Biotechnology 28:723-726.

Genomic epidemiology of novel coronavirus - Global subsampling [Internet]. 2020. Available from: https://nextstrain.org/ncov/global?l=clock

Odon V, Fros JJ, Goonawardane N, Dietrich I, Ibrahim A, Alshaikhahmed K, Nguyen D, Simmonds P. 2019. The role of ZAP and OAS3/RNAseL pathways in the attenuation of an RNA virus with elevated frequencies of CpG and UpA dinucleotides. Nucleic Acids Research 47:8061-8083.

World Health Organization DRAFT landscape of COVID-19 candidate vaccines – 14 July 2020 [Internet]. 2020. Available from: https://www.who.int/publications/m/item/draft-landscape-of-covid-19-candidate-vaccines

Pachetti M, Marini B, Benedetti F, Giudici F, Mauro E, Storici P, Masciovecchio C, Angeletti S, Ciccozzi M, Gallo RC, et al. 2020. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. J Transl Med 18:179.

Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, Harris SR. 2016. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microb Genom 2:e000056.

Peeples L. 2020. News Feature: Avoiding pitfalls in the pursuit of a COVID-19 vaccine. Proceedings of the National Academy of Sciences 117:8218-8221.

Peng G, Lei KJ, Jin W, Greenwell-Wild T, Wahl SM. 2006. Induction of APOBEC3 family proteins, a defensive maneuver underlying interferon-induced anti-HIV-1 activity. The Journal of experimental medicine 203:41-46.

Pruitt KD, Tatusova T, Maglott DR. 2005. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Research 33:D501-504.

Radhakrishnan A, Chen YH, Martin S, Alhusaini N, Green R, Coller J. 2016. The DEAD-Box Protein Dhh1p Couples mRNA Decay and Translation by Monitoring Codon Optimality. Cell 167:122-132 e129.

Rima BK, McFerran NV. 1997. Dinucleotide and stop codon frequencies in single-stranded RNA viruses. Journal of General Virology 78 (Pt 11):2859-2870.

Rocha EP, Smith JM, Hurst LD, Holden MT, Cooper JE, Smith NH, Feil EJ. 2006. Comparisons of dN/dS are time dependent for closely related bacterial genomes. Journal of Theoretical Biology 239:226-235.

Savisaar R, Hurst LD. 2017. Both Maintenance and Avoidance of RNA-Binding Protein Interactions Constrain Coding Sequence Evolution. Molecular Biology and Evolution 34:1110-1126.

Savisaar R, Hurst LD. 2018. Exonic splice regulation imposes strong selection at synonymous sites. Genome Research 28:1442-1454.

Shu Y, McCauley J. 2017. GISAID: Global initiative on sharing all influenza data - from vision to reality. Euro Surveill 22:10.2807/1560-7917.ES.2017.2822.2813.30494.

Simmonds P. 2020. Rampant C->U hypermutation in the genomes of SARS-CoV-2 and other coronaviruses – causes and consequences for their short and long evolutionary trajectories. bioRxiv:2020.2005.2001.072330.

Simmonds P, Xia WJ, Baillie JK, McKinnon K. 2013. Modelling mutational and selection pressures on dinucleotides in eukaryotic phyla -selection against CpG and UpA in cytoplasmically expressed RNA and in RNA viruses. BMC Genomics 14:Artn 610.

Smith EC, Blanc H, Surdel MC, Vignuzzi M, Denison MR. 2013. Coronaviruses lacking exoribonuclease activity are susceptible to lethal mutagenesis: evidence for proofreading and potential therapeutics. PLoS Pathog 9:e1003565.

Stachyra A, Gora-Sochacka A, Sirko A. 2014. DNA vaccines against influenza. Acta Biochimica Polonica 61:515-522.

Stachyra A, Redkiewicz P, Kosson P, Protasiuk A, Gora-Sochacka A, Kudla G, Sirko A. 2016. Codon optimization of antigen coding sequences improves the immune potential of DNA vaccines against avian influenza virus H5N1 in mice and chickens. Virol J 13:143.

Takata MA, Goncalves-Carneiro D, Zang TM, Soll SJ, York A, Blanco-Melo D, Bieniasz PD. 2017. CG dinucleotide suppression enables antiviral defence targeting non-self RNA. Nature 550:124-127.

Thanh Le T, Andreadakis Z, Kumar A, Gomez Roman R, Tollefsen S, Saville M, Mayhew S. 2020. The COVID-19 vaccine development landscape. Nature Reviews: Drug Discovery 19:305-306.

Tulloch F, Atkinson NJ, Evans DJ, Ryan MD, Simmonds P. 2014. RNA virus attenuation by codon pair deoptimisation is an artefact of increases in CpG/UpA dinucleotide frequencies. Elife 3:e04531.

van Dorp L, Acman M, Richard D, Shaw LP, Ford CE, Ormond L, Owen CJ, Pang J, Tan CCS, Boshier FAT, et al. 2020. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infect Genet Evol 83:104351.

Walker PJ, Siddell SG, Lefkowitz EJ, Mushegian AR, Dempsey DM, Dutilh BE, Harrach B, Harrison RL, Hendrickson RC, Junglen S, et al. 2019. Changes to virus taxonomy and the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2019). Archives of Virology 164:2417-2429.

Testing recombination in the pandemic SARS-CoV-2 strains [Internet]. 2020 17.06.2020]. Available from: https://observablehq.com/@spond/linkage-disequilibirum-in-sars-cov-2

Wang S, Taaffe J, Parker C, Solorzano A, Cao H, Garcia-Sastre A, Lu S. 2006. Hemagglutinin (HA) proteins from H1 and H3 serotypes of influenza A viruses require different antigen designs for the induction of optimal protective antibody responses as studied by codon-optimized HA DNA vaccines. Journal of Virology 80:11628-11637.

Wu Q, Medina SG, Kushawah G, DeVore ML, Castellano LA, Hand JM, Wright M, Bazzini AA. 2019. Translation affects mRNA stability in a codon-dependent manner in human cells. Elife 8:10.7554/eLife.45396.

Xia X. 2020. Extreme genomic CpG deficiency in SARS-CoV-2 and evasion of host antiviral defense. Molecular Biology and Evolution:https://doi.org/10.1093/molbev/msaa1094.

Yermilov V, Rubio J, Ohshima H. 1995. Formation of 8-Nitroguanine in DNA Treated with Peroxynitrite in-Vitro and Its Rapid Removal from DNA by Depurination. FEBS Letters 376:207-210.

Zhuang JC, Lin C, Lin D, Wogan GN. 1998. Mutagenesis associated with nitric oxide production in macrophages. Proceedings of the National Academy of Sciences of the United States of America 95:8286-8291.

Figure legends

Figure 1 Chord diagram displaying the rate of flux from one dinucleotide to another in the coding sequence of SARS-CoV-2. For each node, the direction of flux is indicated by the indentation of the connecting links: the outer most layer represents flux into the node and the inner layer represents flux out. The frequency of the flux exchange is represented by the width of any given link where it meets the outer axis. Dinucleotide nodes are coloured according to their GC-content. Hence, it is evident that there is high flux away from GC-rich dinucleotides whereas AU-rich dinucleotides are largely conserved.

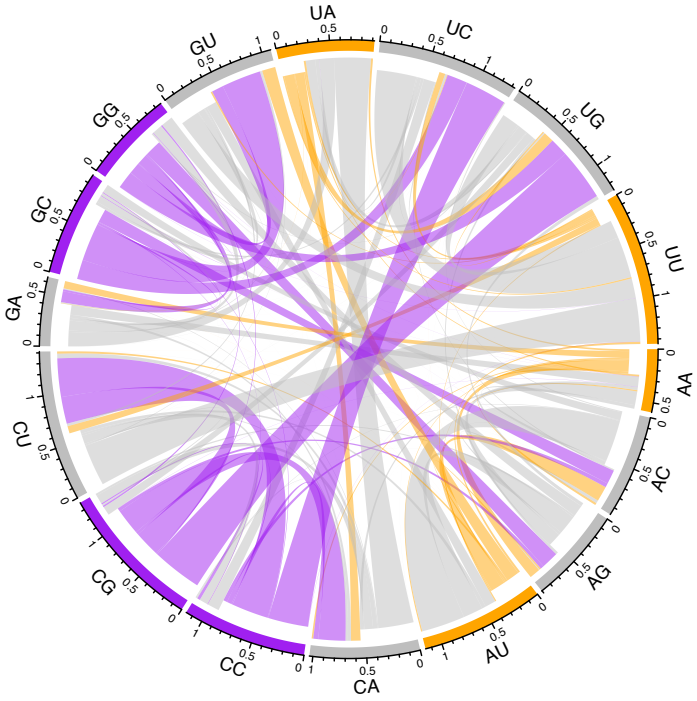


Figure 2. Comparison of dinucleotide content across SARS-CoV-2 compared with neutral expectations. Error bars represent bootstrapped 95% upper and lower confidence bounds.

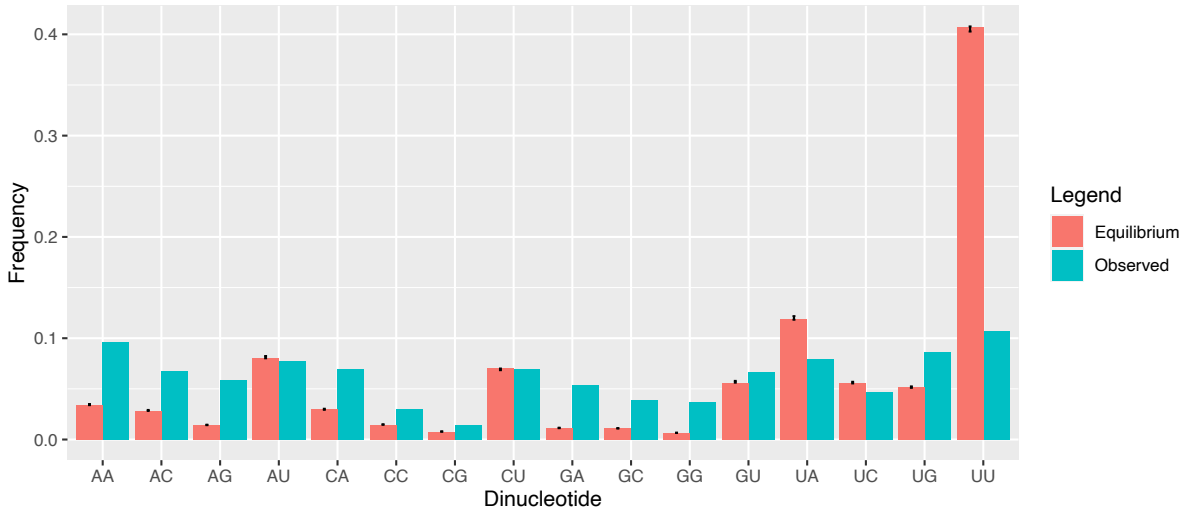


Figure 3. a) CpG enrichment across the genes of SARS-CoV-2. Grey line = no enrichment **b)** relationship between CpG enrichment and GC3

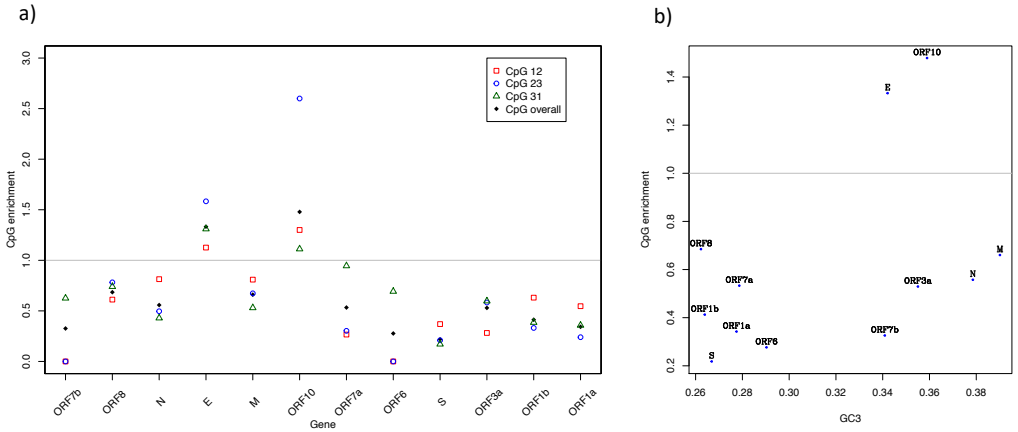


Figure 4. GC content across genes of SARS-CoV-2 at codon sites 1, 2 3 and averaged across the gene

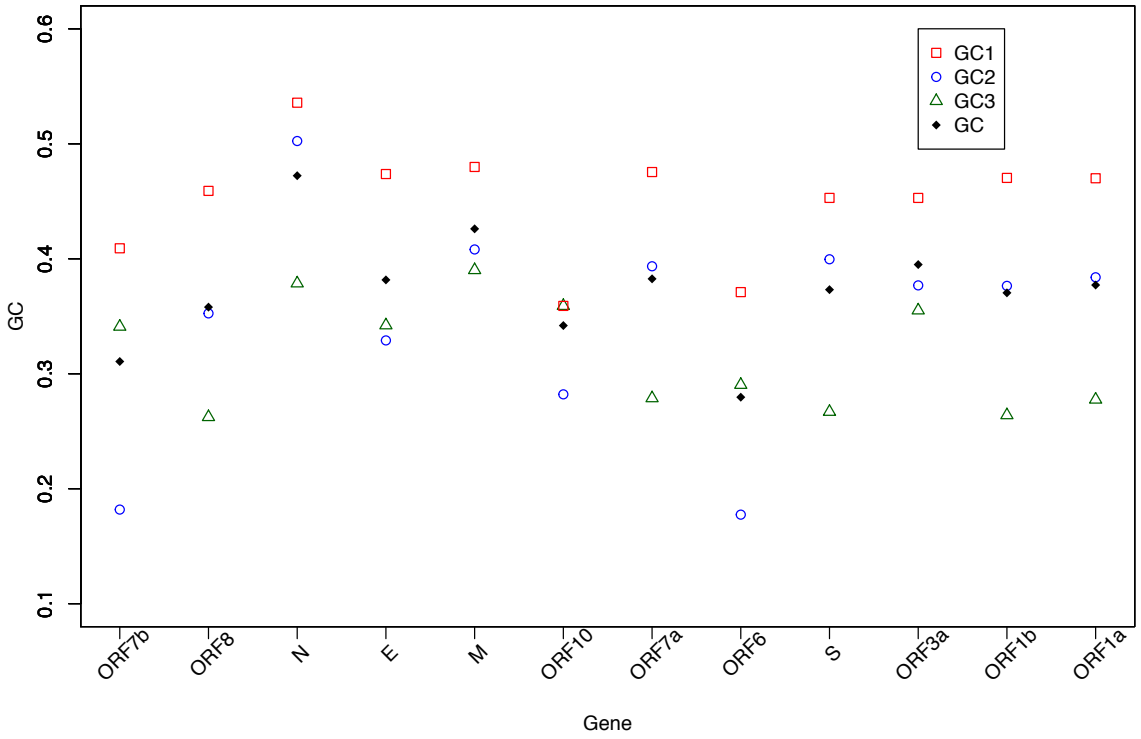


Figure 5. Base composition at codon third sites across genes of SARS-CoV-2.

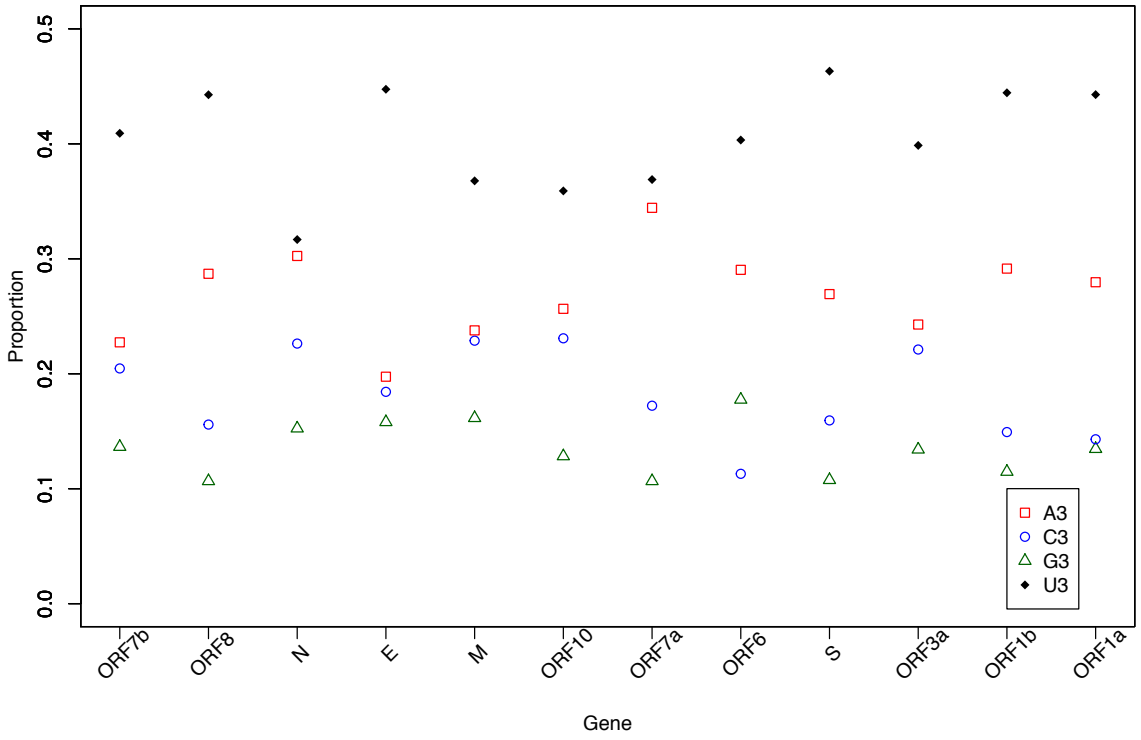


Figure 6. a) UpA enrichment across genes of SARS-CoV-2 **b)** correlation with CpG enrichment. Grey line is the line of slope 1 through the origin.

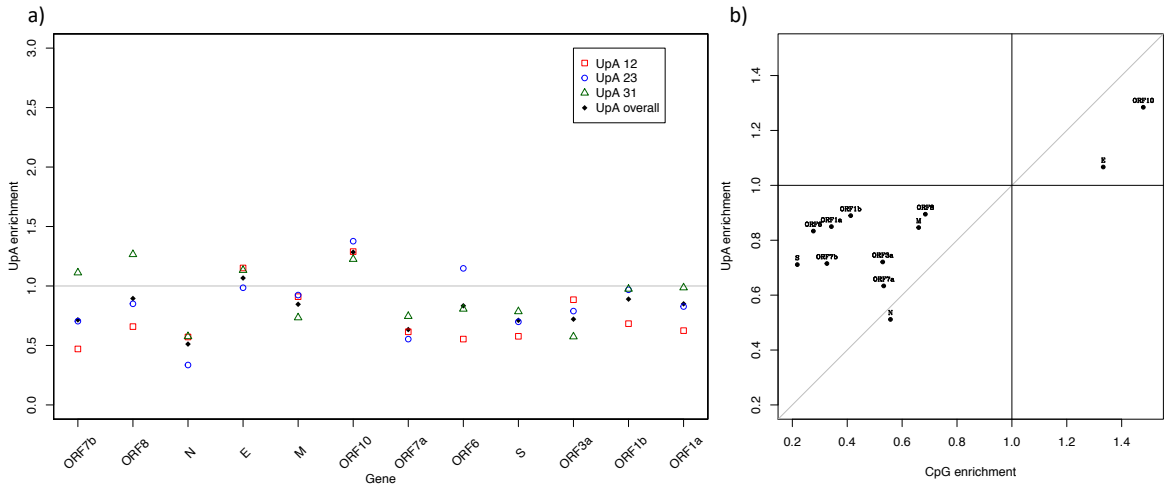


Figure 7 a) GpC and **b)** ApU enrichment across the genes of SARS-CoV-2.

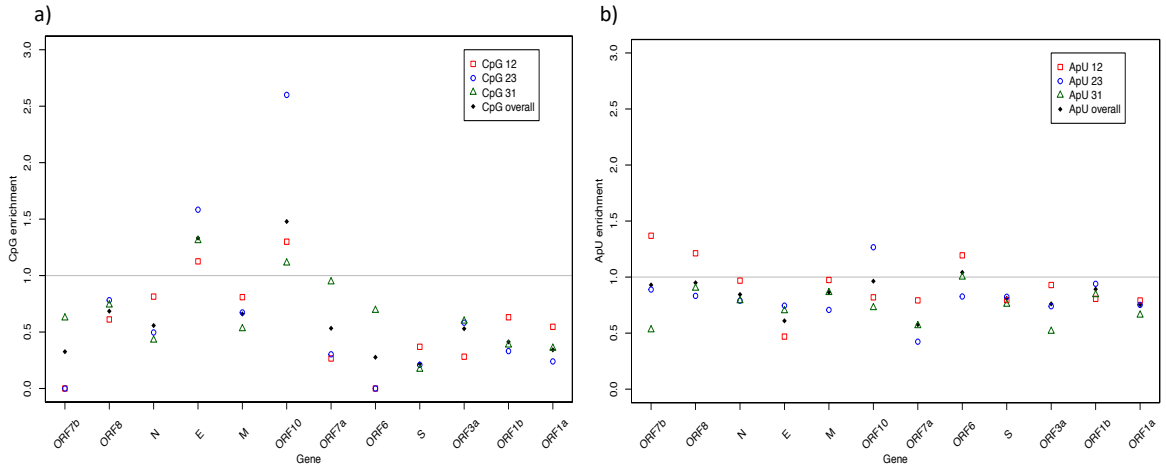


Figure 8 Correlation between expression level and CpG enrichment, GC content and GC3

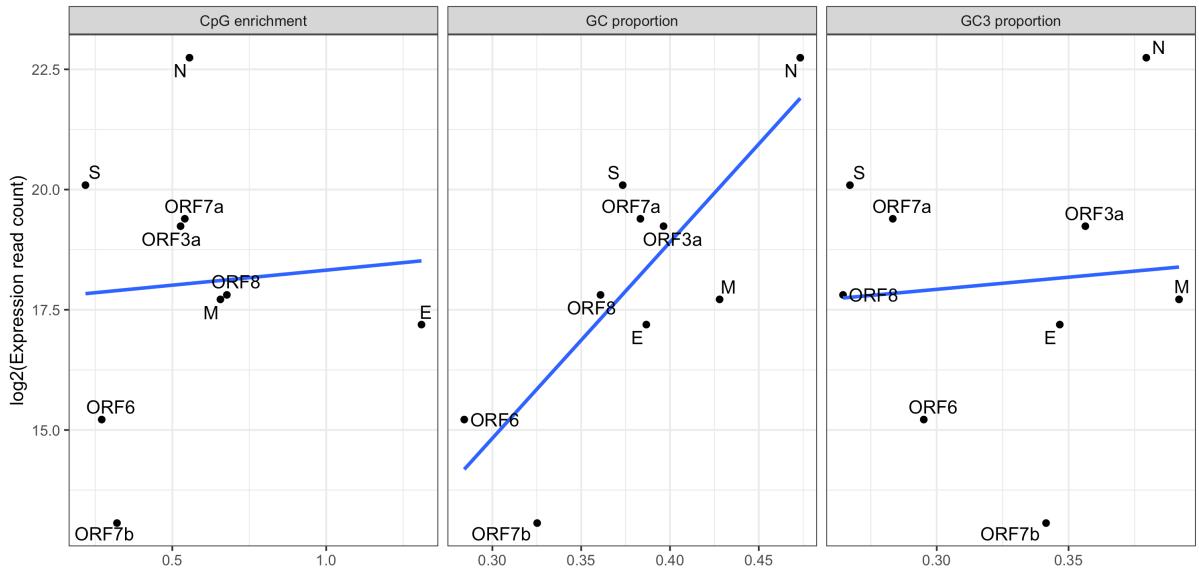
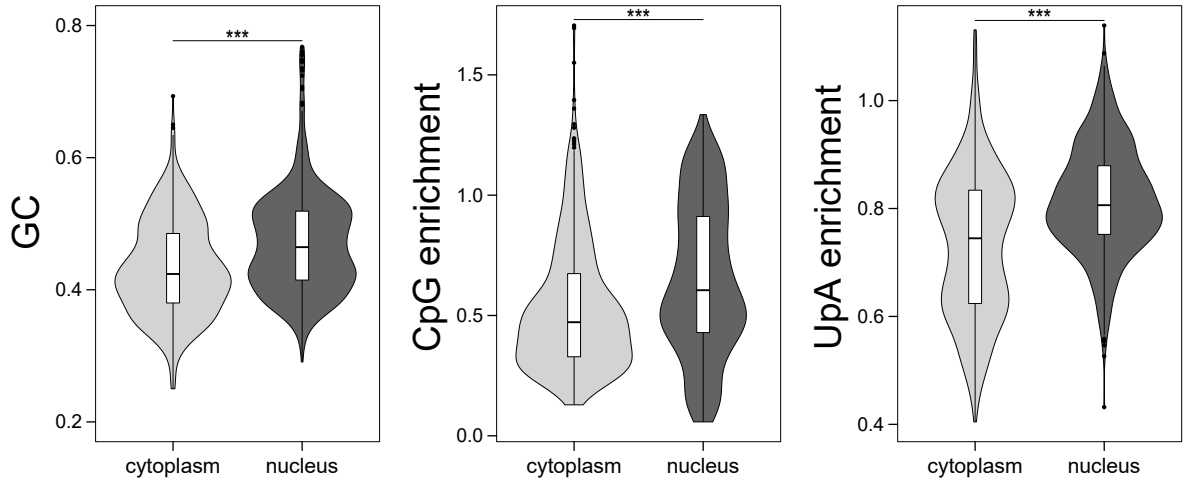


Figure 9. GC content of cytoplasmic and nuclear viruses. Cytoplasmic viruses have significantly lower values for all three measures (MWU test: GC: p = 6.42e-18, CpG enrichment p = 1.35e-13, UpA enrichment: p = 9.1e-29)



Supplementary files:

S Table 1: Ligand binding by antiviral proteins

S Table 2: Data on nucleotide content of viruses

S Table 3 SARS-CoV-2 genomes used and acknowledgement

S Table 4 H1N1 genomes used and acknowledgement

S Table 5 H1N1 genomes used and acknowledgement

S Table 6 The dinucleotide mutational matrix for SARS-CoV-2

S Table 7 The attenuation algorithm and properties of resulting sequences

S Table 8 The proposed attenuated sequences